Application of Gaussian Processes for black-box modelling of biosystems

Abbreviated title suggestion: *GP black-box modelling of biosystems*

Abstract

Different models can be used for nonlinear dynamic system identification and the Gaussian process model is a relatively new option with several interesting features: model predictions contain the measure of confidence, the model has a small number of training parameters and facilitated structure determination, and different possibilities of including prior knowledge exist. In this paper the framework for the identification of a dynamic system model based on the Gaussian processes is shown, illustrated on a simulated bioreactor example and then applied on two case studies. The first one addresses modelling of the nitrification process in a wastewater treatment plant and the second modells the biomass growth in the Lagoon of Venice. Special emphasis is placed on model validation, an often underemphasised part of the identification procedure, where the Gaussian model prediction variance can be utilised.

Key words: nonlinear system identification, Gaussian process model, biotechnological systems

1 Introduction

While there are numerous methods for the identification of linear dynamic systems from measured data, the nonlinear systems identification requires more

sophisticated approaches. The most common choices include *e.g.* artificial neural networks (ANN) and fuzzy models, which can be seen as universal approximators. One of the biggest practical disadvantages ANN has is the curse of dimensionality [1] — the exponential growth of the modelled volume with the input space dimension [2]—, leading to (a) ANN with big number of neurons and (b) a lot of data needed for a system description. The local model network (LMN) [1], a form of the fuzzy model, reduces this problem, but has problems with a description of the off-equilibrium regions of the dynamical system [1,3]. As an alternative, the Gaussian process model was proposed for the identification of nonlinear dynamic systems [3]. Gaussian process (GP) models present a new, emerging, complementary method for nonlinear system identification.

The GP model is a probabilistic, non-parametric black-box model. It differs from most of the other black-box identification approaches as it does not try to approximate the modelled system by fitting the parameters of the selected basis functions but rather searches for the relationship among measured data. Gaussian process models are closely related to approaches such as Support Vector Machines (SVM) and specially Relevance Vector Machines (RVM) [4] due to the use of kernel functions [5]. Also the way RVM reduces the needed parameters is similar to the way GP model "reduces" the number of the equivalent ANN parameters [5]. More on comparison of the GP model and the other methods can be found in [5,6] and references therein.

The output of the Gaussian process model is a normal distribution, expressed in terms of mean and variance. The mean value represents the most likely output and the variance can be interpreted as the measure of its confidence. The obtained variance, which depends on the amount and quality of available identification data, is important information distinguishing the GP models

from other methods. The GP model structure determination is facilitated as only the covariance function and the regressors of the model need to be selected. Another potentially useful attribute of the GP model is the possibility to include various kinds of prior knowledge into the model, see *e.g.* [7] for the incorporation of local models and the static characteristic. Also the number of model parameters, which need to be optimised is smaller than in other black-box identification approaches. The disadvantage of the method is the potential computational burden for optimisation and prediction that increases with amount of data and number of regressors.

The GP model was first used for solving a regression problem in the late seventies, but it gained popularity within the machine learning community in the late nineties of the twentieth century. Results of a possible implementation of the GP model for the identification of dynamic systems were presented only recently, *e.g.* [8,9]. The investigation of the model with uncertain inputs, which enables the propagation of uncertainty through the model, is given in [4,10] and illustrated in [11].

The purpose of this paper is twofold. First, to present the procedure of dynamic system identification using the model based on Gaussian processes. Second, within this framework emphasis is placed on validation, which includes several features [12]: the model purposiveness, the model plausibility and model falseness. Within this two objectives the use of properties differing the GP model from other approaches, *i.e* prediction variance and the way the input/output relations are encompassed, is emphasised. The validation of a dynamic system model based on Gaussian processes is illustrated with a simulated dynamic system example and demonstrated in two biotechnological case studies, already used for the illustration of a GP model validation in [13].

Biotechnological systems are often considered as complex, however simplified input/output behaviour representations are sufficient for certain purposes, *e.g.* feedback control design, prediction models for supervisory control, *etc.* In the paper it is shown how the advantages of Gaussian process models can be used in identification and validation of such models.

The paper is organised as follows. In Section 2 basic principles of the GP model and its use in dynamic system identification are described. The methodology of the identification with a GP model, presented on an illustrative example, is given in Section 3. The identification of two biotechnological systems is given in the next two sections, *i.e.* a wastewater treatment plant in Section 4 and algae growth in the Venice lagoon in Section 5. In the last section the main conclusions are gathered.

2 Modelling of dynamic systems with Gaussian processes

2.1 Modelling with the GP model

Here, modelling with the GP model is presented only in brief, for a more detailed explanation see *e.g.* [6,14].

A Gaussian process is a Gaussian random function, fully described by its mean and variance. Gaussian processes can be viewed as a collection of random variables $f(\mathbf{x}_i)$ with joint multivariate Gaussian distribution: $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n) \sim \mathcal{N}(0, \mathbf{K})$. Elements K_{ij} of the covariance matrix $\mathbf{K}$ are covariances between values of the function $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ and are functions of corresponding arguments $\mathbf{x}_i$ and $\mathbf{x}_j$: $K_{ij} = C(\mathbf{x}_i, \mathbf{x}_j)$. Any function $C(\mathbf{x}_i, \mathbf{x}_j)$ can be a covariance

function, providing it generates a nonnegative definitive covariance matrix $\mathbf{K}$.

Certain assumptions about the process are made implicitly with the covariance function selection. The *stationarity of the process* results in the value of covariance function $C(\mathbf{x}_i, \mathbf{x}_j)$ between inputs $\mathbf{x}_i$ and $\mathbf{x}_j$ depending only on their normalised Euclidian distance [1] and being invariant to their translation in the input space, see *e.g.* [6]. *Smoothness of the output* reflects in outputs $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ having higher covariance when inputs $\mathbf{x}_i$ and $\mathbf{x}_j$ are closer together. The common choice [6,14] for the covariance function, representing these assumptions, is the Gaussian covariance function:

$$C(\mathbf{x}_i, \mathbf{x}_j) \;=\; \mathrm{Cov}[f(\mathbf{x}_i), f(\mathbf{x}_j)] \;=\; v \exp\left[-\frac{1}{2}\sum_{d=1}^{D} w_d (x_i^d - x_j^d)^2\right] + \delta_{ij} v_0 \quad (1)$$

where D is the dimension of the input space of vector $\mathbf{x}$ and $\boldsymbol{\Theta} = [w_1, \ldots, w_D, v, v_0]^T$ is a vector of parameters called hyperparameters [2]. Additional linear terms of the covariance function $+a_0 + a_1 \sum_{d=1}^{D} (x_i^d x_j^d)$ from [14] are omitted as stationarity of the process is presumed. Hyperparameter v controls the magnitude of the covariance and hyperparameters w_i represent the relative importance of each component x^d of vector $\mathbf{x}$. The part $\delta_{ij} v_0$ represents the covariance between outputs due to white noise [3], where δ_{ij} is the Kronecker operator and v_0 is the white noise variance. With the use of covariance function (1) the total number of the GP model parameters is $D+2$ for the size D input, where for example the number of comparable artificial neural networks parameters would be, due to the curse of dimensionality problem, considerably larger.

[1] also Mahalanobis distance [15].

[2] The hyperparameters are higher level parameters of the model which include noise variance and regularisation constants [16].

[3] When assuming different kinds of noise the covariance function should be changed appropriately, *e.g.* [16].

Apart from the parameter v_0, which expresses the process noise variance, no physical insight about the underlying system is given by the hyperparameters, which makes a GP model a black-box model.

The GP model fits nicely into the Bayesian modelling framework. The idea behind GP modelling is to place the prior directly over the space of functions instead of parameterizing the unknown function $f(\mathbf{x})$ [6]. The simplest type of such a prior is Gaussian. Consider the system

$$y(k) = f(\mathbf{x}(k)) + \epsilon(k) \tag{2}$$

with white Gaussian noise $\epsilon(k) \sim \mathcal{N}(0, v_0)$ with variance v_0 and the vector of regressors $\mathbf{x}(k)$ from operating space $\mathcal{R}^D$. We put the GP prior with covariance function (1) with unknown hyperparameters on the space of functions $f(.)$.

Within this framework we have $y_1, \ldots, y_N \sim \mathcal{N}(0, \mathbf{K})$ with $\mathbf{K} = \mathbf{\Sigma} + v_0\mathbf{I}$, where $\mathbf{\Sigma}$ is the covariance matrix for the noise-free system (2) and $\mathbf{I}$ is $N \times N$ identity matrix. Based on a set of N training data pairs $\{\mathbf{x}_i, y_i\}_{i=1}^N$ we wish to find the predictive distribution of y_{N+1} corresponding to a new given input $\mathbf{x}_{N+1}$. For the collection of random variables $(y_1, \ldots, y_N, y_{N+1})$ we can write:

$$\begin{pmatrix} \mathbf{y} \\ \\ y_{N+1} \end{pmatrix} \sim \mathcal{N}(0, \mathbf{K}_{N+1}) \tag{3}$$

with covariance matrix

$$\mathbf{K}_{N+1} = \begin{bmatrix} \begin{bmatrix} \\ \mathbf{K} \\ \\ \end{bmatrix} & \begin{bmatrix} \\ \mathbf{k}(\mathbf{x}_{N+1}) \\ \\ \end{bmatrix} \\ \\ \begin{bmatrix} \mathbf{k}(\mathbf{x}_{N+1})^T \end{bmatrix} & \begin{bmatrix} k(\mathbf{x}_{N+1}) \end{bmatrix} \end{bmatrix} \tag{4}$$

where $\mathbf{y} = [y_1, \ldots, y_N]^T$ is an $N \times 1$ vector of training targets, $\mathbf{k}(\mathbf{x}_{N+1}) = [C(\mathbf{x}_1, \mathbf{x}_{N+1}), \ldots, C(\mathbf{x}_N, \mathbf{x}_{N+1})]^T$ is the $N \times 1$ vector of covariances between training inputs and the test input and $k(\mathbf{x}_{N+1}) = C(\mathbf{x}_{N+1}, \mathbf{x}_{N+1})$ is the autocovariance of the test input. We can divide this joint probability into a marginal and a conditional part. The marginal term gives us the likelihood of the training data: $p(\mathbf{y}|\mathbf{X}) \sim \mathcal{N}(0, \mathbf{K})$, where $\mathbf{X}$ is the $N \times D$ matrix of training inputs.

We need to estimate the unknown hyperparameters $\boldsymbol{\Theta} = [w_1, \ldots, w_D,\ v,\ v_0]^T$ of the covariance function (1), where the parameter v_0 is the estimate of the noise by which the training data is corrupted. This is usually done via maximization of the log-likelihood

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Theta}) &= \log(p(\mathbf{y}|\mathbf{X})) = \\ &= -\frac{1}{2}\log(\mid \mathbf{K} \mid) - \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} - \frac{N}{2}\log(2\pi) \end{aligned} \tag{5}$$

with the vector of hyperparameters $\boldsymbol{\Theta}$ and $N \times N$ training covariance matrix $\mathbf{K}$, where the hyperparameters distribution $p(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{X})$ is approximated with their most likely values. The optimization requires the computation of the

derivative of $\mathcal{L}$ with respect to each of the parameters:

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \Theta_i} = -\frac{1}{2}\text{trace}\left(\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \Theta_i}\right) + \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \Theta_i}\mathbf{K}^{-1}\mathbf{y} \tag{6}$$

Here, it involves the computation of the inverse of the $N \times N$ covariance matrix $\mathbf{K}$ at every iteration, which can be computationally demanding for large N. Second option is to approximate the hyperparameters distribution $p(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{X})$ using Markov Chain Monte Carlo (MCMC) methods. The reader is referred to *e.g.* [6] for detailed description of parameter optimisation methods.

Given that the hyperparameters are known, we can obtain a prediction of the GP model at the input $\mathbf{x}_{N+1}$. The conditional part of (3) provides the predictive distribution of y_{N+1}:

$$p(y_{N+1}|\mathbf{y}, \mathbf{X}, \mathbf{x}_{N+1}) = \frac{p(\mathbf{y}, y_{N+1})}{p(\mathbf{y}|\mathbf{X})} \tag{7}$$

It can be shown [14] that this distribution is Gaussian with mean and variance:

$$\mu(\mathbf{x}_{N+1}) = \mathbf{k}(\mathbf{x}_{N+1})^T\mathbf{K}^{-1}\mathbf{y} \tag{8}$$
$$\sigma^2(\mathbf{x}_{N+1}) = k(\mathbf{x}_{N+1}) - \mathbf{k}(\mathbf{x}_{N+1})^T\mathbf{K}^{-1}\mathbf{k}(\mathbf{x}_{N+1}). \tag{9}$$

Vector $\mathbf{k}(\mathbf{x}_{N+1})^T\ \mathbf{K}^{-1}$ in (8) can be interpreted as a vector of smoothing terms which weights training outputs $\mathbf{y}$ to make a prediction at the test point $\mathbf{x}_{N+1}$. If the new input is far away from the data points, the term $\mathbf{k}(\mathbf{x}_{N+1})^T\ \mathbf{K}^{-1}\ \mathbf{k}(\mathbf{x}_{N+1})$ in (9) will be small, so that the predicted variance $\sigma^2(\mathbf{x}_{N+1})$ will be large. Regions of the input space, where there are few data or are corrupted with noise, are in this way indicated through higher variance. From the system identification point of view the equation (8) provides the model prediction and equation (9) its confidence.

The presented GP model was originally used for modelling static nonlinearities, but it can be extended to model dynamic systems as well [4,8,10]. Our task is to model the dynamic system (2), where

$$\mathbf{x} = [y(k-1), \ldots, y(k-L),\ u(k-1), \ldots, u(k-L)] \tag{10}$$

is the vector of regressors that determines nonlinear ARX model structure of the L-th order, and be able to make multi-step ahead model prediction.

One way to do multi-step ahead prediction is to make iterative one-step ahead predictions up to desired step whilst feeding back the predicted output. Two general approaches to iterated one-step ahead prediction are possible using the GP model. In the first only the mean values of the predicted output are fed back to the input. In this, so called "naive" approach, the input vector $\mathbf{x}$ into the GP model at time step k is:

$$\mathbf{x} = [\hat{y}(k-1), \ldots, \hat{y}(k-L),\ u(k-1), \ldots, u(k-L)] \tag{11}$$

In the second, so called "exact", approach the complete output distributions are fed back. This way the information in predicted variance is not lost and the model simulation gives more accurate results, especially in terms of predicted variance, which is not overconfident as in "naive" approach. More on the GP model simulation and differences of approaches can be found *e.g.* in [8,10].

3 Gaussian process model identification methodology

In this section the framework for dynamic system identification with GP models is given. The identification framework consists of roughly six stages:

(1) defining the purpose of the model,
(2) model selection,
(3) design of the experiment,
(4) realisation of the experiment and data processing,
(5) training of the model and
(6) model validation.

The model identification is an iterative process. Returning to some previous procedure step is possible at any step in the identification process and is usually necessary.

3.1 The model purpose and model selection

The decision for the use of a specific model derives from the model purpose and from the limitations met at the identification process. In this paper selection of the GP model is presumed. This approach can be beneficial when the information about the system exists in the form of input/output data, when data are corrupted, *e.g.* by noise and measurement errors, when a measure of confidence in model prediction is required and when there is a relatively small amount of data in respect to the selected number of regressors.

After the model is selected, its structure must be determined next. In the case of the GP model this means selecting the covariance function and the model

regressors. The choice of the covariance function reflects the relationship between data and is based on prior knowledge of the process. The standard choice for smooth and stationary processes is function (1). Prior knowledge about other attributes, *e.g.* periodicity, non-stationarity, can be expressed through a different choice of the covariance function [6].

The second part of structure determination is the choice of suitable regressors. In the case of a dynamic system model this also means selecting the model order, which is the area of intensive research, as it is common difficulty to all nonlinear identification methods. It has been pointed out in [17], that the nonminimal realisation of the model might be required to capture the dynamic of the nonlinear system, in accordance with Taken's embedding theorem [18], which determines the the necessary order of the model obtained from sampled input-output data. For the selection of the GP model input space the reader is also referred to [19]. A review of the general approaches for the choice of regressors is given in [20].

The most frequent approach for regressor selection is the so called *validation based regressor selection* [20], where the search for the optimal vector of regressors is initiated from some basic set of regressors. After the model optimisation and cross-validation, the regressors are added to or taken from the model. Prospering models according to selected performance are kept while dissatisfying models are rejected. In the case of normalised inputs the influence of each regressor can be observed through the value of the associated hyperparameter. If the associated regressor is not relevant enough it can be removed from the perspective model.

3.2 Obtaining data – design of the experiment, experiment and data processing

Data describing the unknown system is very important in any black-box identification. For a good description of the process the influential variables and suitable sample time must be chosen.

The design of the experiment and the experiment itself are, as is always the case in systems modelling, very important parts of the identification procedure. The quality of the model depends on the system information contained in the measurement data, regardless of the identification method. Nevertheless, the design of the experiment is not the focus of this paper. More information on this topic can be found *e.g.* in [21].

As already mentioned the Gaussian process modelling approach relies on the relation among input/output data and not on approximation with basis functions. Consequently, this means that the distribution of identification data within the process operating region is crucial for the quality of the model. Model predictions can be informative only if the inputs to the model lie in the regions, where training data is available. The GP model is good for interpolation, but not for extrapolation, which is indicated by large variances of model predictions.

Consequently, the data for model training should be chosen reasonably, which can be obstructed by the nature of the process (*e.g.* limitations in the experiment design in industrial processes, physical limitations of the system). To cancel the influence of different measuring scales, the preprocessing of measured data can be pursued, *e.g* centering and scaling, in this paper referred as

normalisation.

3.3 Model training

In the GP model approach training means optimization of hyperparameters $\boldsymbol{\Theta}$ from (1). Each hyperparameter w_d expresses the relative importance of the associated regressor, similar to the automatic relevant detection (ARD) method [14], where a higher value of w_d expresses higher importance of the regressor. Hyperparameter v expresses the overall scale of correlations and hyperparamter v_0 accounts for the influence of noise [14]. Several possibilities of hyperparameter determination exist. A very rare possibility is that hyperparameters are known in advance as prior knowledge. Almost always, however, they must be determined from the training data, where different approaches are possible, *e.g.* [10]. Mostly the likelihood maximization (ML) approach is used as it gives good results despite its simplification, where any optimization method could be used to achieve ML [10].

3.4 Model validation

Validation concerns the level of agreement between the mathematical model and the system under investigation [22] and it is many times underemphasised despite its importance. Several features can represent the quality of the model. Their overview can be found *e.g.* in [12,22]. The most important are model plausibility, model falseness and model purposiveness, explained as follows.

Model *plausibility* expresses the model's conformity with the prior process knowledge by answering two questions: whether the model "looks logical" and

whether the model "behaves logical". The first question addresses the model structure, which in the case of GP models means mainly the plausibility of the hyperparameters. The second one is concerned with the responses of the model output to typical events on the input, which can be validated with visual inspection of the responses as is the case with other black-box models.

Model *falseness* reflects the agreement between the process and the model output or the process input and the output of the inverse model. The comparison can be done in two ways, both applicable to GP models: qualitatively, *i.e.* by visual inspection of differences in responses between the model and the process, or quantitatively, *i.e.* through evaluation of performance measures. Beside commonly used performance measures such as *e.g.* mean relative square error (MRSE, [23]), which compares only the mean prediction of the model to the output of the process:

$$\mathrm{MRSE} = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} y_i^2}} \tag{12}$$

where y_i and $e_i = \hat{y}_i - y_i$ are the system's output and prediction error in i-th step of simulation, the performance measures such as log predictive density error (LD, [8,10]) can be used for evaluating GP models, taking into account not only mean prediction but the entire predicted distribution:

$$\mathrm{LD} = \frac{1}{2}\, log(2\pi) + \frac{1}{2N} \sum_{i=1}^{N} \left(log(\sigma_i^2) + \frac{e_i^2}{\sigma_i^2} \right) \tag{13}$$

where σ_i^2 is the prediction variance in i-th step of simulation. Performance measure LD weights the prediction error e_i more heavily when it is accompanied with smaller predicted variance σ_i^2, thus penalising overconfident predictions more than acknowledged bad predictions, indicated by higher variance. Another possible performance measure, applicable in the training procedure,

is the negative log-likelihood of the training data (LL, [10]):

$$\mathrm{LL} = \frac{1}{2}\log \mid \mathbf{K} \mid + \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} + \frac{N}{2}\log(2\pi), \tag{14}$$

where $\mathbf{K}$ is the covariance matrix, $\mathbf{y}$ is the vector of targets and N is the number of training points. LL is the measure inherent to the hyperparameter optimisation process, see (5), and gives the likelihood that the training data is generated by given, *i.e.* trained, model. The smaller the MRSE, LD and LL are, the better the model is.

Variance of the model predictions on a validation signal can be a validation measure itself, as it indicates whether the model operates in the region, where identification data were available. Nevertheless, it should be used carefully and in combinations with other validation tools, as predictions with small variance are not necessary good, as it will be shown in the following illustrative example.

Model *purposiveness* or usefulness tells whether or not the model satisfies its purpose, which means the model is validated when the problem that motivated the modelling exercise can be solved using the obtained model. Here, again, the prediction variance can be used, *e.g.* when the prediction confidence is too low, the model can be labeled as not purposive.

3.5 An illustrative example of GP model dynamic system identification

The purpose of this example is to demonstrate the GP model identification procedure with special emphasis on the utility of prediction variance and other GP model specific measures for model validation. The example illustrates how validation is used as the selection criteria for the best model. The selected model is then used to demonstrate the influence of increased noise variance

on the system's output, the behaviour of the model prediction in unmodelled regions and the behaviour of the model when a new, unmodelled input is introduced to the system.

The second order discrete bioreactor model [24] was taken as the system to be identified for demonstration purposes. In the bioreactor the microorganisms grow by consuming the substrate. The bioreactor is given as the discrete second order dynamical system [24] with sampling time $T_s = 0.5$s:

$$\begin{aligned} x_1(k+1) &= x_1(k) + 0.5\frac{x_1(k)x_2(k)}{x_1(k)+x_2(k)} - 0.5u(k)x_1(k) \\ x_2(k+1) &= x_2(k) - 0.5\frac{x_1(k)x_2(k)}{x_1(k)+x_2(k)} - 0.5u(k)x_2(k) + 0.05u(k) \\ y(k) &= x_1(k) + \epsilon(k) \end{aligned} \tag{15}$$

where $x_1(k)$ is the concentration of the microorganisms and $x_2(k)$ is the concentration of the substrate. The control input $u(k)$ is the output flow rate, limited between $0 \leq u(k) \leq 1$. Output of the system $y(k)$ is the concentration of microorganisms, corrupted by white Gaussian noise $\epsilon(k)$ with standard deviation $\sigma = 0.0005$. Our task is to model this system with the GP model and validate the acquired model.

To acquire the identification data, system (15) was excited with the control input signal u in the form of 4s long stairs with random amplitude values between $0 \leq u(k) \leq 0.7$. Note that the upper limit of the input signal was chosen so that a part of the operating region remained unmodelled. The bioreactor was modelled with fourth, third and second order model. Before training of the models the signals were normalised, so that they had a maximum value of one and a minimum value of minus one. From normalised signals 602 training points were composed. i-th training point at the sample step k for the L-th

order GP model, $L = 2, 3, 4$, is composed from the input regressors:

$$\mathbf{x}_i = [y_n(k-1), \ldots, y_n(k-L),\ u_n(k-1), \ldots, u_n(k-L)]$$

and the output value $y_i = y_n(k)$, where u_n and y_n are normalised input and output signals. Gaussian covariance function (1) was chosen for the covariance function as the smooth and stationary output is presumed and considered as a prior.

All three GP models were validated with simulation, where the validation data was obtained by simulating the system (15) using similar but not the same control input signal $u(k)$ as for obtaining identification data.

The results of the regressor selection procedure can be seen in Table 1. Performance measure LL was used with the training data and performance measures MRSE and LD were used on the validation data.

From the performance measures used on the validation results, shown in the first three rows of Table 1, can be seen that the second order model proves best among three identified models in terms of MRSE and LD values of the simulation results. Beside the MRSE and LD values of the simulation results, the second order model is also favoured by the Occam's razor principle [6], stating that the most simple explanation of the given problem should be used.

Hyperparameters w_{x^i} reflect the relative importance of regressors $x(k-i)$ and in all model structures the regressor $y(k-1)$ can be excluded due to the small value of the associated hyperparameter w_{y^1}. The removal of this regressor from the selected second order model results in even better validation results.

This regressor selection procedure lead us to the GP model, in Table 1 marked

as ♠, with the following regressors: $[y(k-2), u(k-1), u(k-2)]$.

[Insert Table 1 about here]

The simulation results on the validation data for the selected second order model can be seen in Figure 1, where the model's output and the noise free target are depicted. It can be seen that most of the time the value of the predicted standard deviation σ is around $5{\cdot}10^{-4}$, corresponding to the noise level present at the system's output. The prediction variance increases at the steps where the input u changes its value due to the small number of training points describing those regions, resulting in the increase of the $k(\mathbf{x}) - \mathbf{k}(\mathbf{x})\mathbf{K}^{-1}\mathbf{k}(\mathbf{x})$ part of the (9). It can also be observed that the error of the model prediction remains inside the 95% confidence limits, defined within $\pm 2\sigma_i$, indicating the level of trust which can be put in the prediction.

[Insert Figure 1 about here]

These model validation results will serve as the reference for the observation of how different conditions can influence the model prediction and validation.

First, it will be shown how model prediction changes when the model reaches the unmodelled region of the system. As there is no training data available nearby, the model must extrapolate from the data describing the neighbouring regions in order to make predictions. This worsens the prediction mean, but is also accompanied by the increase of the prediction variance, thus widening the model confidence limits. This effect can be observed in Figure 2, where the values of the control input were increased above the $u(k) > 0.7$ at time $t > 12$s.

[Insert Figure 2 about here]

Second, we would like to show how the increase of the system's output noise variance reflects in the identified model. For this purpose the standard deviation of the system's output noise was increased to $\sigma = 2{\cdot}10^{-3}$. The control input signal, used for generating identification and validation data, was the same as in the reference example. The second order GP model is identified and again the regressor $y(k-1)$ is removed. The values of the GP model hyperparameters can be seen in Table 1 (the line, marked with ♣).

The mean model prediction is satisfactory and the prediction variance is increased on account that higher output noise variance is predicted, as can be observed from the simulation results on validation data in Figure 3. The estimation of the system's output noise is satisfactory close to the real value, *i.e.* $\sigma = 2.1{\cdot}10^{-3}$, which also shows, that the value of the hyperparameter v_0 tends to the value of the system's output noise when enough training data is used. The value of MRSE is slightly worse as in the reference example, as this model is identified with more noise present in the training data. Also the value of LD is slightly worse as, despite the increased variance, the influence of the prediction error prevails.

[Insert Figure 3 about here]

Finally, we would like to show how the unmodelled regressor influences the model. For this purpose an extra regressor in the form of additional control input z, not correlated to input u, was added to the system. The effect of this input is the same as the effect of the control input u and could represent an

additional outlet or leak of the system (15), which changes the description to:

$$\begin{aligned} x_1(k+1) &= x_1(k) + 0.5\frac{x_1(k)x_2(k)}{x_1(k)+x_2(k)} - 0.5u(k)x_1(k) - 0.5z(k)x_1(k) \\ x_2(k+1) &= x_2(k) - 0.5\frac{x_1(k)x_2(k)}{x_1(k)+x_2(k)} - 0.5u(k)x_2(k) \\ &\quad -0.5z(k)x_2(k) + 0.05u(k) + 0.05z(k) \\ y(k) &= x_1(k) + \epsilon(k) \end{aligned} \tag{16}$$

The reference GP model was used for prediction, where the input z is not present and therefore neglected at the training of the model. The control input in the form of a step $z = 0.05$ is introduced into the system at validation time 30s. The (non)influence of the unmodelled regressor on the prediction variance, when the model operates in the region with sufficient training data, can be seen in Figure 4. The model prediction from time $t = 30$s worsens, but the 95% confidence limits remain tight. This example shows that the variance cannot be informative on the unaccounted influences on the system in the identification data.

[Insert Figure 4 about here]

With the bioreactor example the following properties of the GP model have been illustrated:

(1) The hyperparameters' ARD property can be effectively used to reduce the number of regressors of the identified model.

(2) There are two possible causes for the increase of the prediction variance:
   - the particular region of the system, where the model makes predictions, is described with insufficient training data and
   - the data, describing particular regions, contains more noise. In the ex-

ample this has been shown for the whole region, but the same goes when noise is increased only in part of the system's operating region.

These two causes can not be easily distinguished without prior knowledge about the identified system.

(3) When nonmodelled influence is introduced to the system, the model prediction, including the variance, does not change.

4 Case study 1 – Activated sludge model for a wastewater treatment plant

4.1 Introduction

In the first case study we would like to show the possibilities and attributes of the GP model identification and associated model validation by means of nitrification process identification in a wastewater treatment plant (WWTP). The process of interest is moving bed biofilm reactor pilot plant in Domžale-Kamnik WWTP. The case study is adopted from [23], where it was identified with linear and parametric nonlinear models.

The pilot plant (see Figure 5) consists of an anoxic reactor ($176m^3$), two aerobic reactors (130 and $117m^3$), a mixed reactor ($115m^3$) and a settler ($600m^3$) [23,25]. The main process of ammonia (NH_4-N) removal is nitrification, which runs in aerobic reactors. It is strongly dependant on dissolved oxygen (DO) concentrations in the aerobic reactors. The other process variables that influence the nitrification process are dissolved influent flowrate, influent ammonia concentration and temperature. The identified variable is the ammonia concentration at the outlet. The biomass in the plant is attached to plastic

carriers, which flow freely in the anoxic and both aerobic reactors. High DO concentration is needed to drive the diffused oxygen into the biofilm.

[Insert Figure 5 about here]

4.2 Model purpose

The purpose of the developed model is multi-step ahead prediction, which could be used *e.g.* for model predictive control (MPC) or for model based supervisory control.

4.3 Identification data

For the identification the same data was used as in [23]. We did not have the possibility to influence the data acquisition. This also slightly changes the order of the identification procedure steps. The data describing the WWTP was gathered during several weeks of experiments in February and April 2004. The measured quantities were the ammonia concentration at the inlet S_{in} and at the outlet S_{out} of the plant, the input flowrate Φ, the DO concentrations in both aerobic reactors DO_1, DO_2 and the temperature T. Sample time of the measurements was $T_s = 15$ minutes. As the input flowrate was almost constant its influence on the output of the process was neglected. The measured data can be seen in Figure 6 with February data (Set 1) on the left and April data (Set 2) on the right. Due to the damaged sensor the values of DO_2 concentration for the first seven days in Set 1 are false and were not used.

[Insert Figure 6 about here]

Data from both sets was divided into training and validation data sets. The samples from the beginning and the end of the data sequences were used for training and the five-day intervals in the middle of the data sequences were used for validation, as seen in Figure 7, resulting in training and validation data consisting of data from both sets. The purpose of such selections lies in the nature of the GP model, which demands the training data distributed in all regions, where the predictions are conceived. Before training the data was normalised, so that all the signals had a maximum value of one and a minimum value of minus one. Altogether about 1200 input/output samples were used for the training of the GP model.

[Insert Figure 7 about here]

4.4 Model structure selection and model training

As a stationary, nonlinear process with smooth output was presumed, the covariance function (1) is chosen. The regressors were selected based on the *validation based regressor selection procedure*, the method explained in the illustrative example, but with modification due to the large number of the potential regressors and limited amount of the identification data. The easiest way to select the regressors would be to consider all the potentially useful regressors at once, train the model and then use the ARD property of the hyperparameters to eliminate the regressors with low impact. The problem is that the system has a lot of control inputs and the number of potentially

needed regressors is too large for the available amount of training data [4]. Therefore the regressor selection procedure was broken down to smaller steps, where particular regressor groups were tested and the resulting models validated. Models giving better results were kept, their regressors not carrying enough information neglected and new, potentially beneficial regressors were added. This procedure is longer and more difficult to pursue, as in principle many possible regressor combinations in the regressor pool must be tested.

Based on the prior process knowledge from [23] and its authors, the following regressors were chosen for the initial GP model structure: $[S_{out}^{1,2,3}\ S_{in}^{4}\ T^{2}\ DO_{1}^{3}\ DO_{2}^{1}]$, where the superscript denotes the delay, *e.g.* $S_{in}^{4} = S_{in}(k-4)$. For the validation and comparison of different identified models the same performance measures were used as in illustrative example, *i.e.* negative log-likelihood of identification data (14), mean relative square error (12), log-predictive density (13) and visual inspection of the validation data. Values of the performance measures for only some among many of the identified GP models can be seen in Table 2 to illustrate the main search direction among the different sets of regressors. The ARD property of the GP models' hyperparameters was used to remove the regressors not contributing much information.

[Insert Table 2 about here]

Through this procedure a satisfactory GP model, marked with † in Table 2,

[4] Even if there had been enough training data, the training of the GP model of such dimensions would have demanded considerable amounts of computer time due to the inversion of the large covariance matrix, see (6), at every optimisation step [6].

with regressors $[S_{out}^{1,2,3}\ S_{in}^{3,4}\ T^{2,6}\ DO_1^{2,3}\ DO_2^{2,4}]$ was obtained. The hyperparameters of this model are given in Table 3. We can observe, that three of the hyperparameters w_d are relatively small compared to others, *i.e.* the hyperparameters corresponding to the regressors $S_{out}^2,\ S_{in}^4$ and T^2, smaller than 10^{-4}. The three regressors were removed, leading to the final model, in Table 2 marked with ♠. Because of removing of some of the less informative regressors the value of LL on identification data increased, but according to MRSE and LD values on validation data, the model performs best among all of the tested GP models.

[Insert Table 3 about here]

4.5 Model validation

The "naive" simulation (Section 2.2) with validation data Sets 1 and 2 (Figure 7) was selected for the qualitative validation of the GP model. The results of the simulation can be seen in Figure 8. The peaks and crashes of both the model and the process responses coincide, which suggests satisfactory dynamic behaviour, despite noticeable error in some regions. Relatively low estimated system's output noise variance $v_0 = 0.0191$ and the tight prediction confidence limits relatively to the prediction error indicate that some of the system functional dependencies remained unmodelled, as demonstrated in the previous illustrative example.

[Insert Figure 8 about here]

Comparison of the model performance with that in [23], where the model has

similar dynamic behaviour and MRSE around 0.45, indicates that the acquired GP model is slightly better. However, it should be noticed here that the division of the data on training and validation set was different in [23], where the entire Set 2 was used for training and the entire Set 1 for validation. For the first principles modelling approach the selection of training data is not that crucial, as the structure of the parametric model reflects the assumed underlying physical and chemical principles among the data in the entire operating region, while for the GP model the relationship between the data is not explicit and the model must extrapolate in the regions, where no data is available for modelling.

The model in [23] was identified for the design of a model predictive controller. To validate the GP model usefulness, the model was used for the multi-step ahead prediction. Dependance of the MRSE and LD for the multi-step ahead prediction up to the horizon $k = 24$ (6h) are depicted in Figure 9. The results of a 24-step ahead prediction, just as the simulation results, suggest unmodelled system dependencies through the high value of the LD measure and very tight confidence limits. Still, the mean prediction, evaluated with the MRSE is satisfactory.

[Insert Figure 9 about here]

As the GP model performs slightly better than the models identified in [23] and labeled as useful, we can conclude that the identified GP model would be *useful* too, despite being *false* and *not plausible* in some regions.

5 Case study 2 – Lagoon of Venice

The purpose of this example is solely to demonstrate a possible utility of confidence measure provided with the Gaussian process model. The development of the input/output model for biomass growth in the Lagoon of Venice is taken as a case study. The Lagoon of Venice measures 550 km^2, but is very shallow, with an average depth of less than 1 m. It is heavily influenced by anthropogenic inflow of nutrients 7 mio kg/year of nitrogen and 1.4 mio kg/year of phosphorus [26]. These (mainly nitrogen) loads are above the Lagoons admissible trophic limit and generate its dystrophic behaviour, characterised by excessive growth of algae, mainly *Ulva rigida*. Very modest sets of measured data were available [27] for modelling the growth of algae in the Lagoon. The data were sampled weekly for slightly more than one year. The sampled quantities are nitrogen in ammonia NH_3, nitrogen in nitrate NO_3, total nitrogen N, phosphorus in orthophosphate PO_4, all in [μg/l], dissolved oxygen DO in percentage of saturation, temperature T [°C], and algae biomass B, dry weight in [g/m^2].

The purpose of the model is the prediction of the algae growth in the lagoon, making the algae biomass B the output of the system. Other sampled quantities are taken as the possible inputs. As the amount of available training data is very small, only the first order model with seven regressors is presumed.

Two sets of measurements with 43 sampled values were utilised, both heavily corrupted with the measurement errors of order 20-50% [26]: one for the GP model training and other for the model validation. The normalisation of the measurement data had been used before identification.

As in previous examples, the validation based regressor selection method was used for the selection of regressors. The lack of the identification data in this study is even more stressed than in the previous example. Several combinations of regressors for the first order model structures were tried and the models were validated with the visual inspection and the use of performance measures MRSE (12) and LD (13) of the one-step ahead prediction results on validation data. Due to the scarce information contained in the identification data there was no significant difference between identified models, making regressor selection hard to pursue.

Through the validation based regressor selection method the following input regressors were selected for the prediction of algae biomass $B(k)$ at the time step k:

$$[B(k-1)\ T(k-1)\ DO(k-1)\ N(k-1)],$$

with the values of corresponding hyperparameters:
$w_1 = 177,\ w_2 = 6.6{\cdot}10^{-7},\ w_3 = 32,\ w_4 = 5,\ v = 2.8$ and hyperparameter $v_0 = 1.0{\cdot}10^{-3}$, corresponding to the estimate of noise standard deviation of $\sigma = 33.8$. The regressor $T(k-1)$ was removed from the model due to the small value of hyperparameter w_2, resulting in the model with following regressors:

$$[B(k-1)\ DO(k-1)\ N(k-1)].$$

The results of the one-step ahead prediction on the identification and validation data can be seen in Figure 10 and the results of the simulation on the validation data set can be seen in Figure 11. The values of the performance measures for one-step ahead prediction with the identification data and one-step ahead prediction and simulation with training data are given in Table 4.

[Insert Figure 10 about here]

[Insert Figure 11 about here]

[Insert Table 4 about here]

The model fits the identification data well, but the presented one-step ahead prediction and simulation results on validation data exhibit high MRSE. Nevertheless, the results are comparable to those obtained by other modelling investigations of the same case study, *e.g.* [26] [5] and references in there, where they are evaluated as potentially acceptable if peaks and crashes of the biomass concentration are predicted correctly.

However, the 95% confidence limits in Figures 10 and 11, depicting the model's one-step ahead and simulation results on the validation data set clearly indicate that the confidence in the obtained model is very low [6]. From the prediction variance it can be concluded that this GP model is not to be trusted, regardless of the potentially acceptable mean values of the predictions, as it predicts in the region where not enough training information was available. This decision can be made straightforwardly based on the available confidence limits, which come as a handy validation utility. Also the relatively low

[5] In [26] the combination of knowledge-driven and data-driven approach to modelling was used. In one case the inputs DO and NH_3 are used to predict the state B, in the second case the input T is used additionally.

[6] Note that for demonstration purposes no limitations have been induced on the prediction, resulting in mean values and the 95% confidence limits sometimes sinking below the physically impossible concentration zero.

estimate of noise standard deviation $\sigma = 33.8$ suggest, that more training data should be used, as the predicted variance does not match the estimated measurement errors like in *e.g* the bioreactor example, where the increase of system's output noise variance was accompanied with corresponding increase of the prediction variance.

This example was used to demonstrate the use of prediction variance for the validation of the identified GP model. The model itself could be improved with the use of more identification data.

6 Conclusion

In this paper the Gaussian process model is used for dynamic systems identification with emphasis on some of its properties: model predictions containing the measure of confidence, low number of parameters and facilitated structure determination.

The GP model identification procedure with emphasis on the validation has been illustrated with a simulated example and applied to two case studies. The wastewater treatment plant case study resulted in a purposeful model, while the algae growth model of the second case study showed that the lack of information content in training data prevents the development of a trustful Gaussian process model.

The prediction variance is one of the main differences between the GP model and other models. It can be effectively used in the usefulness validation, where the lack of confidence in the model prediction can serve as the grounds to reject the model as not useful. The prediction variance can also be used in falseness

validation, whether via specific performance measures such as log-predictive density error, or through observation of confidence limits around the predicted output. Despite its usefulness in model validation, it should be accompanied with standard validation tools, as the small variance does not necessarily mean that the model is of good quality, as shown in the case study one.

In the validation based regressor selection procedure the log-predictive density error and the log-likelihood of the training data can be useful in selecting model regressors as shown in the first case study. In the case of normalised inputs, the model hyperparameters indicate the influence of corresponding regressors and can be used as a tool for removal of uninfluental regressors at the regressor selection stage of the model selection.

Small amount of data relative to the number of selected regressors, data corrupted with noise and measurement errors and the need for the measure of model prediction confidence could be the reasons to select identification with the GP model. If there is not enough data or it is heavily corrupted with noise, even the GP model can not perform well, but in that case the inadequacy of the model and the identification data is indicated through higher variance of the predictions as seen in case study two.

As the presented results have shown the GP model's potential for the identification of nonlinear dynamic systems, the next step would be to apply the GP model to problems, where the advantages of the GP model could be effectively used, *e.g.* control design, diagnostic system design *etc.*

Acknowledgment

The authors gratefully acknowledge the contribution of the Ministry of Higher Education, Science and Technology of Republic of Slovenia, Grant No. P2-0001.

References

[1] Johansen, T.A., Shorten, R. and Murray-Smith, R., On the interpretation and identification of dynamic Takagi-Sugeno fuzzy models. IEEE Transactions on Fuzzy Systems, 2000, **8**(3), 297–313.

[2] Murray-Smith, R. and Johansen, T.A., Multiple model approaches to modelling and control. Taylor and Francis, London, 1997.

[3] Murray-Smith, R., Johansen, T.A. and Shorten, R., On transient dynamics, off-equilibrium behaviour and identification in blended multiple model structures. In: Proc. in European Control Conference, Karlsruhe, Germany, 1999, BA-14.

[4] Quiñonero-Candela, J. and Rasmussen, C.E., Analysis of Some Methods for Reduced Rank Gaussian Process Regression. In: Murray-Smith, R. and Shorten R. (Eds.), Switching and Learning in Feedback Systems, Lecture Notes in Computer Science, Vol. 3355, 2005.

[5] Bishop, C.M. and Tipping, M.E., Bayesian regression and classification. In: Suykens, J., Horvath, G., Basu, S., Micchelli, C. and Vandewalle, J. (Eds.), Advances in Learning Theory: Methods, Models and Applications, NATO Science Series III: Computer and Systems Sciences, IOS press, Amsterdam, 2003, **190**.

[6] Rasmussen, C.E. and Williams, C.K.I., Gaussian Processes for machine learning. The MIT Press, Cambridge, MA, 2006.

[7] Solak, E., Murray-Smith, R., Leithead, W.E., Leith, D.J. and Rassmusen, C.E., Derivative observations in Gaussian Process Models of Dynamic Systems. In: Becker, S., Thrun, S. and Obermayer, K. (Eds.), Advances in Neural Information Processing Systems 15, MIT Press, 2003, 529–536.

[8] Kocijan, J., Girard, A., Banko, B. and Murray-Smith, R., Dynamic Systems Identification with Gaussian Processes. Mathematical and Computer Modelling of Dynamical Systems, 2005, **11**, 411–424.

[9] Gregorčič, G. and Lightbody, G., Gaussian Processes for Modelling of Dynamic Non-linear Systems. In: Proc. Irish Signals and Systems Conference, Cork, Ireland, 2002.

[10] Girard, A., Approximate Methods for Propagation of Uncertainty with Gaussian Process Models. PhD thesis, University of Glasgow, Glasgow, UK, 2004.

[11] Ažman, K. and Kocijan, J., An example of Gaussian process model identification. In: Budin, L. and Ribarić S. (Eds.), Proc. 28th International conference MIPRO – CIS , Opatija, Croatia, 2005, 79–84.

[12] Hvala, N., Strmčnik, S., Šel, D., Milanič, S. and Banko, B., Influence of model validation on proper selection of process models — an industrial case study. Computers and Chemical Engineering, 2005, **29**, 1507–1522.

[13] Ažman, K. and Kocijan, J., Gaussian process model validation: biotechnological case studies. In: Troch, I. and Breitenecker, F. (Eds.), Proc. of the 5th Vienna Symposium on Mathematical Modeling – MathMod, Vienna, Austria, 2006.

[14] Williams C.K.I., Prediction with Gaussian processes: from linear regression and beyond. In: Jordan, M.I. (Ed.), Learning and Inference in Graphical models,

Kluwer Academic Press, 1998, 599–621.

[15] Principe, J.C., Euliano, N.R. and Lefebvre, W.C., Neural and adaptive systems – fundamentals through simulations. John Wiley and Sons, New York, 2000.

[16] Gibbs, M.N., Bayesian Gaussian processes for regression and classification. PhD thesis, Cambridge University, Cambridge, UK, 1997.

[17] Leith, D.J., Leithead, W.E., Solak, E. and Murray-Smith, R., Divide and conquer identification using Gaussian processes. In: Proc. of 41st Conference on Decision and Control, Las Vegas, AZ, 2002, 624–629.

[18] Stark, J., Broomhead, D.S., Davies, M.E. and Huke, J., Delay embeddings of forced systems: II Stochastic forcing. Journal of Nonlinear Science, 2003, **13**(6), 519-577.

[19] Gregorčič, G., Data-based modelling of nonlinear systems for control. PhD thesis, University College Cork, National University of Ireland, Cork, Ireland, 2004.

[20] Lind, I., Regressor selection in system identification using ANOVA. PhD thesis, University of Linköping, Linköping, Sweden, 2001.

[21] Ljung, L., System identification – Theory for the user, 2nd edition. Prentice Hall, New Jersey, 1999.

[22] Murray-Smith, D.J., Methods for the external validation of continuous system simulation models: a review. Mathematical and Computer Modelling of Dynamical Systems, 1998, **4**, 5–31.

[23] Stare, A., Hvala, N. and Vrečko, D., Modelling, identification and validation of models for predictive ammonia control in a wastewater treatment plant - a case study. ISA Transactions, 2006, **45**(2), 159–174.

[24] Cho, J., Principe, J.C., Erdogmus, D. and Motter, M.A., Quasi-sliding model control strategy based on multiple-linear models. Neurocomputing, 2007, **70**, 960–974.

[25] Olsson, G. and Newell, B., Wastewater Treatment Systems. Modelling, Diagnosis and Control. IWA Publishing, London, UK, 1999.

[26] Todorovski, L. and Džeroski, S., Integrating Knowledge-Driven and Data-Driven Approaches to Modelling. Ecological Modelling, 2006, **194**, 3–13.

[27] Coffaro, G., Carrer, G. and Bendoricchio, G., Model for Ulva Rigida growth in the Lagoon of Venice. Technical report, UNESCO MURST Project: Venice Lagoon Ecosystem, University of Padova, Padova, Italy, 1993.

Table 1

Values of validation performance measures and hyperparameters of different bioreactor GP models

Model	Id. data	Valid. data		Hyperparameters									noise‡
order	LL*	MRSE*	LD*	w_{y^4}	w_{y^3}	w_{y^2}	w_{y^1}	w_{u^4}	w_{u^3}	w_{u^2}	w_{u^1}	v	σ
4	-1628	$8.5{\cdot}10^{-3}$	-6.11	$1.2{\cdot}10^{-2}$	$2.6{\cdot}10^{-2}$	$4.3{\cdot}10^{-2}$	$2{\cdot}10^{-6}$	$6{\cdot}10^{-6}$	$1.7{\cdot}10^{-2}$	$1.8{\cdot}10^{-2}$	$1.8{\cdot}10^{-2}$	6.0	$5.0{\cdot}10^{-4}$
3	-1621	$7.5{\cdot}10^{-3}$	-6.12	×	$2.5{\cdot}10^{-3}$	$6.6{\cdot}10^{-2}$	$4{\cdot}10^{-4}$	×	$3.6{\cdot}10^{-3}$	$3.6{\cdot}10^{-3}$	$3.6{\cdot}10^{-3}$	18	$5.2{\cdot}10^{-4}$
2	-1612	$6.1{\cdot}10^{-3}$	-6.41	×	×	$3.0{\cdot}10^{-2}$	$1.1{\cdot}10^{-3}$	×	×	$5.2{\cdot}10^{-3}$	$5.1{\cdot}10^{-3}$	28	$5.3{\cdot}10^{-4}$
2 ♠	-1612	$4.4{\cdot}10^{-3}$	-6.42	×	×	$3.0{\cdot}10^{-2}$	×	×	×	$5.2{\cdot}10^{-3}$	$5.1{\cdot}10^{-3}$	28	$5.3{\cdot}10^{-4}$
2 ♣	-793	$5.0{\cdot}10^{-3}$	-5.19	×	×	$4.3{\cdot}10^{-2}$	×	×	×	$9.2{\cdot}10^{-3}$	$9.2{\cdot}10^{-3}$	10.4	$2.1{\cdot}10^{-3}$

Notes:

‡ identified noise standard deviation σ

♠ reduced number of regressors by neglecting the regressor $y(k-1)$

♣ identified on the output signal with increased noise standard deviation, $\sigma = 0.002$

Table 2

Identification and validation of the GP models with various structures — the values of the performance measures

Regressors					Ident. data	Validation data*	
S_{out}	S_{in}	T	DO_1	DO_2	LL	MRSE	LD
3,2,1	4	2	2	1	-3590	1.950	148
3,2,1	4	4	2,4	1,5	-3617	0.828	271
3,2,1	3,4	2,4	2,4	1,2	-3643	0.325	114
3,2,1	3,4,7	2,4	2,4	1,2,4	-3649	0.321	114
† 3,2,1	3,4	2,6	2,3	2,4	-3649	0.313	108
♠ 3,1	**3**	**6**	**2,3**	**2,4**	**-3640**	**0.295**	**96**
3,2,1	3,4,7	2,4,6	2,3,4	2,4	-3650	0.316	110
3,2,1	3,4,7	2,4,6	2,4,6	1,2,4	-3650	0.319	112
3,2,1	3,4,9	2,6,9	2,3	1,2,4	-3657	0.318	113

Notes: * Validation data Sets 1 and 2 combined

† Selected GP model

♠ Selected GP model after removing uninformative regressors

Table 3

Values of the hyperparameters for the selected GP model

	S_{out}			S_{in}		T		DO_1		DO_2		v	v_0
delay	1	2	3	3	4	2	6	2	3	2	4		
model 1†	0.089	$1.02{\cdot}10^{-6}$	$1.45{\cdot}10^{-2}$	$3.0{\cdot}10^{-3}$	$8.3{\cdot}10^{-5}$	$4.4{\cdot}10^{-5}$	$2.1{\cdot}10^{-4}$	$4.1{\cdot}10^{-4}$	$1.10{\cdot}10^{-3}$	$1.33{\cdot}10^{-3}$	$3.3{\cdot}10^{-4}$	6.2	$1.33{\cdot}10^{-4}$
model 2♠	0.2	×	$2.7{\cdot}10^{-2}$	$2.9{\cdot}10^{-3}$	×	×	$1.07{\cdot}10^{-2}$	$6.8{\cdot}10^{-4}$	$1.27{\cdot}10^{-3}$	$3.4{\cdot}10^{-3}$	$8.8{\cdot}10^{-4}$	2.6	$1.32{\cdot}10^{-4}$

Notes: The values of hyperparameters for normalised data

† Selected GP model

♠ Selected GP model after removing uninformative regressors

Table 4

The values of the performance measures for one-step ahead prediction and simulation results

	Ident. data	Valid. data	
	one-step	one-step	sim
MRSE	0.068	1.13	1.10
LD	4.7	8.3	7.4

Figure captions:

Figure 1: Validation with simulation for the bioreactor GP model, $|e|$ is the absolute value of the error between the predicted and the true values

Figure 2: GP model prediction in the not modelled region

Figure 3: Influence of the increased system's output noise variance on the GP model

Figure 4: Influence of the nonmodelled input on the GP model prediction

Figure 5: The scheme of the Domžale-Kamnik wastewater treatment plant

Figure 6: Process data for the identification, Set 1 from February (left) and Set 2 from April 2004 (right)

Figure 7: Dividing identification data on training and validation data for Set 1 (left) and Set 2 (right), shown for the identified variable S_{out}

Figure 8: Validation with simulation for Set 1 (left) and Set 2 (right)

Figure 9: Values of the MRSE (left) and LD (right), depending on the length of horizon

Figure 10: Comparison of one-step ahead predictions for identification (left) and validation data (right)

Figure 11: Simulation of the GP model with validation data

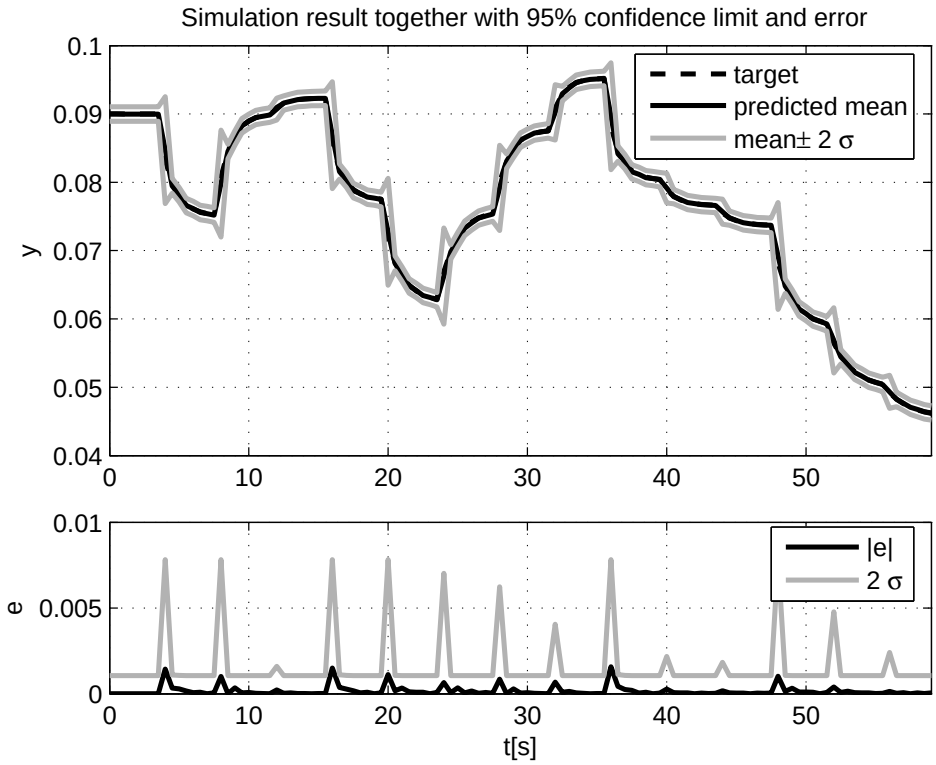
Simulation result together with 95% confidence limit and error
target
predicted mean
mean± 2 σ
y
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0
10
20
30
40
50
|e|
2 σ
e
0.01
0.005
0
t[s]

Fig. 1.

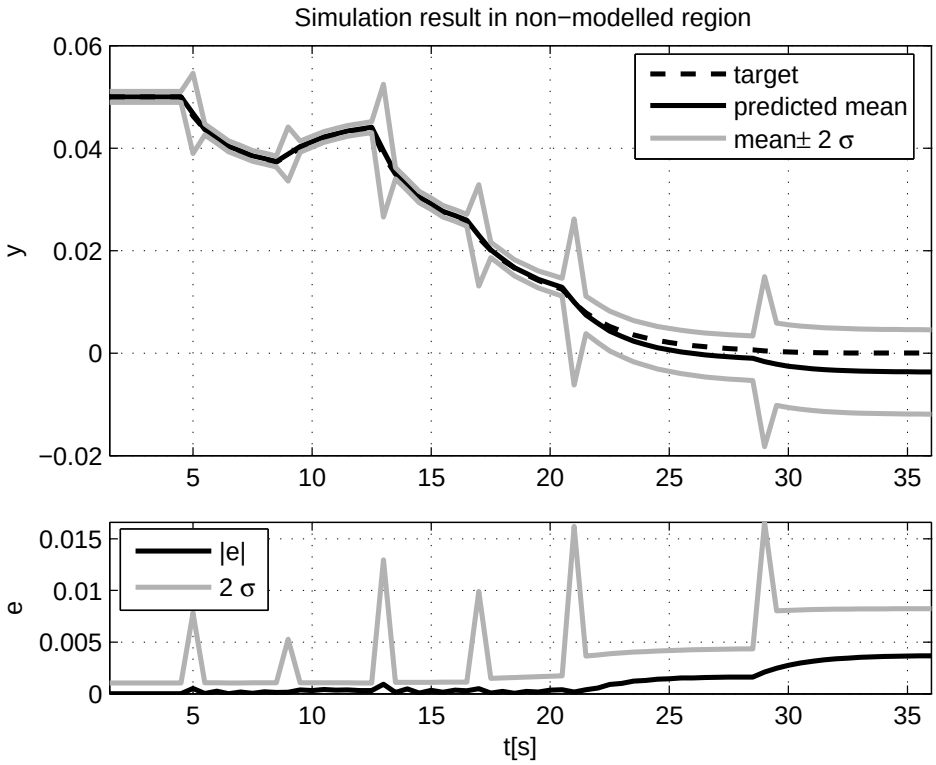
Simulation result in non−modelled region
target
predicted mean
mean± 2 σ
y
0.06
0.04
0.02
0
−0.02
5
10
15
20
25
30
35
|e|
2 σ
e
0.015
0.01
0.005
0
t[s]

Fig. 2.

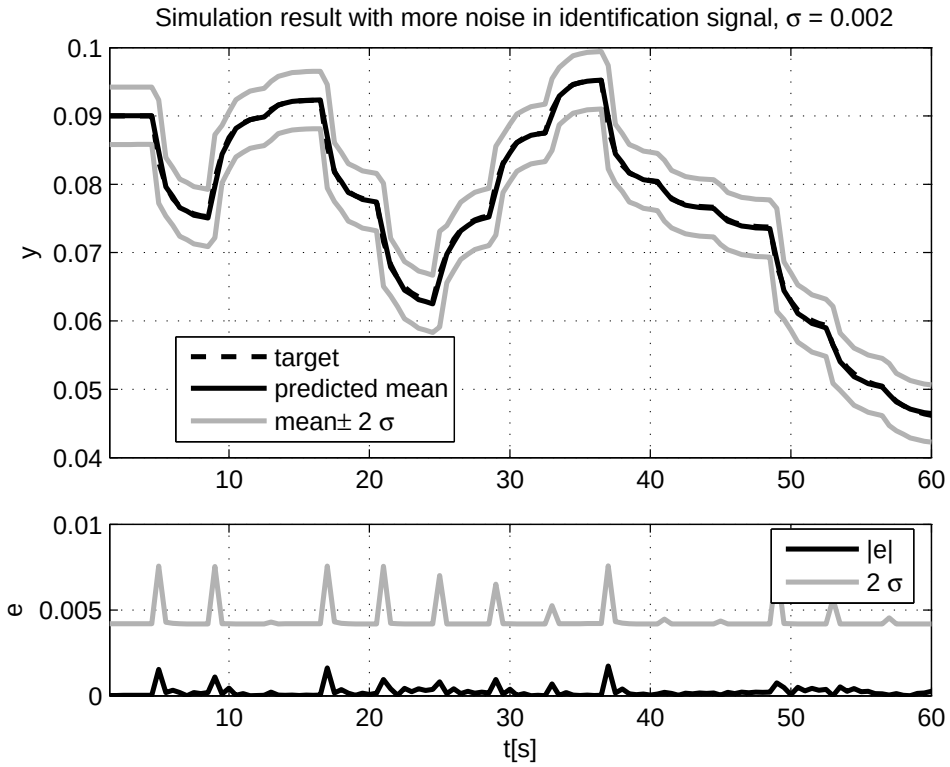
Simulation result with more noise in identification signal, σ = 0.002
y
target
predicted mean
mean± 2 σ
e
|e|
2 σ
t[s]

Fig. 3.

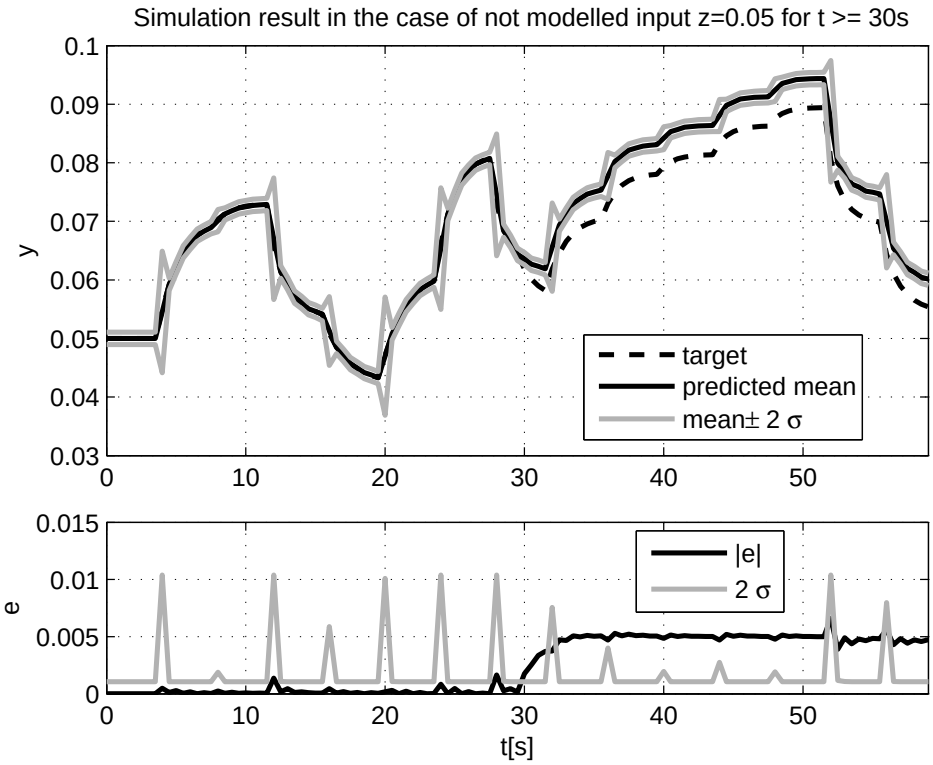
Simulation result in the case of not modelled input z=0.05 for t >= 30s
y
target
predicted mean
mean± 2 σ
e
|e|
2 σ
t[s]

Fig. 4.

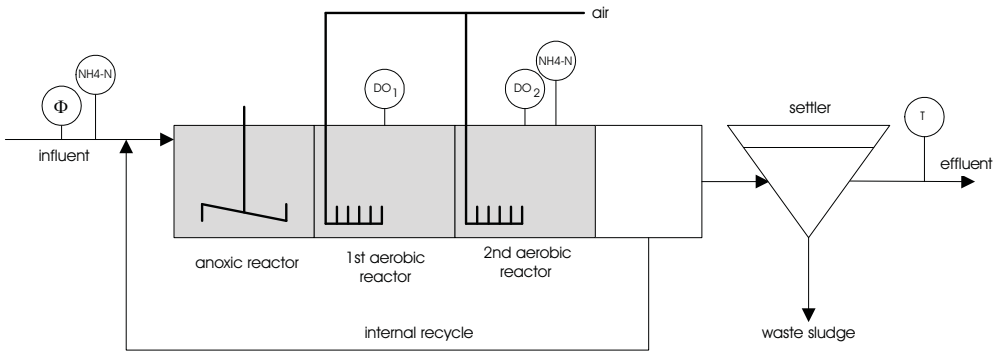
air
NH4-N
Φ
influent
DO 1
DO 2
NH4-N
settler
T
effluent
anoxic reactor
1st aerobic reactor
2nd aerobic reactor
internal recycle
waste sludge

Fig. 5.

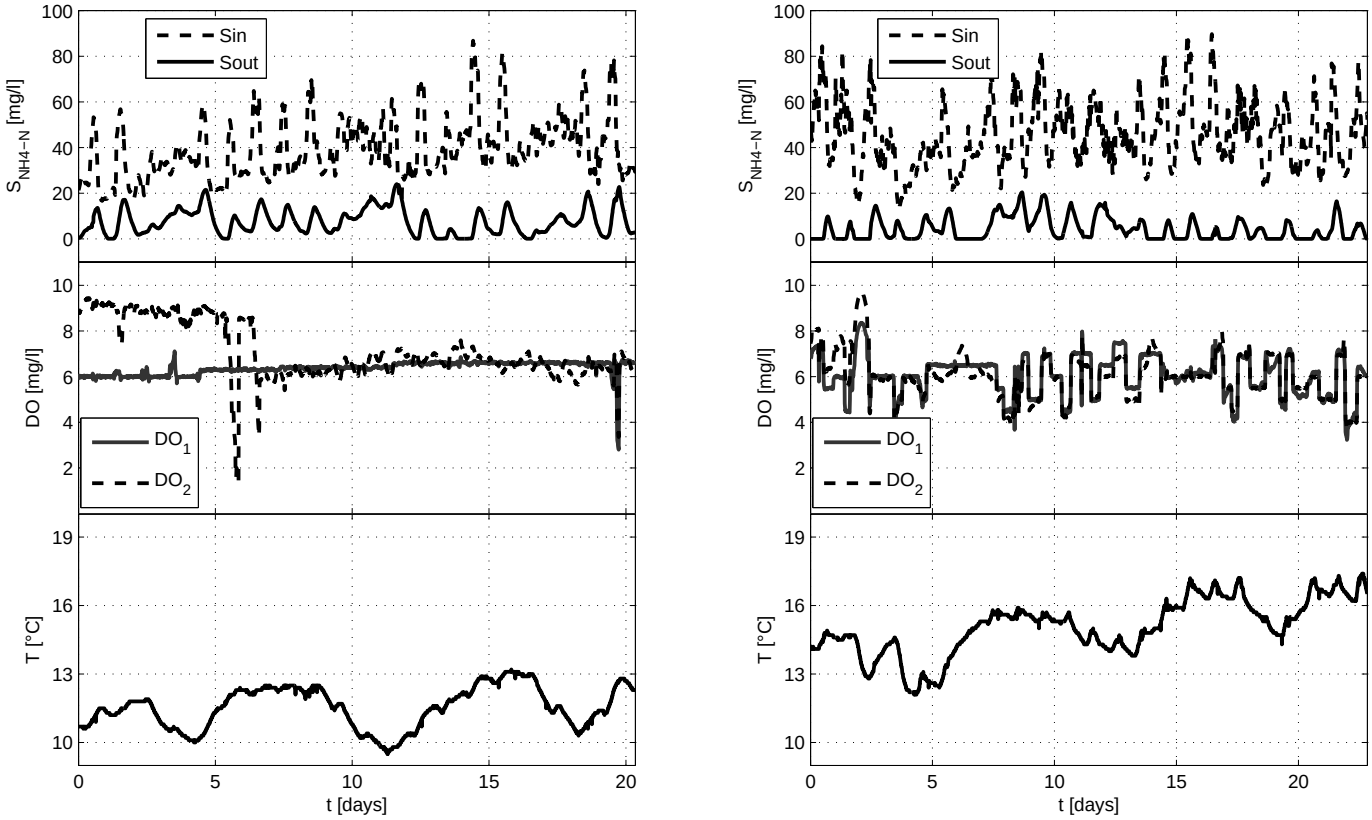

Fig. 6.

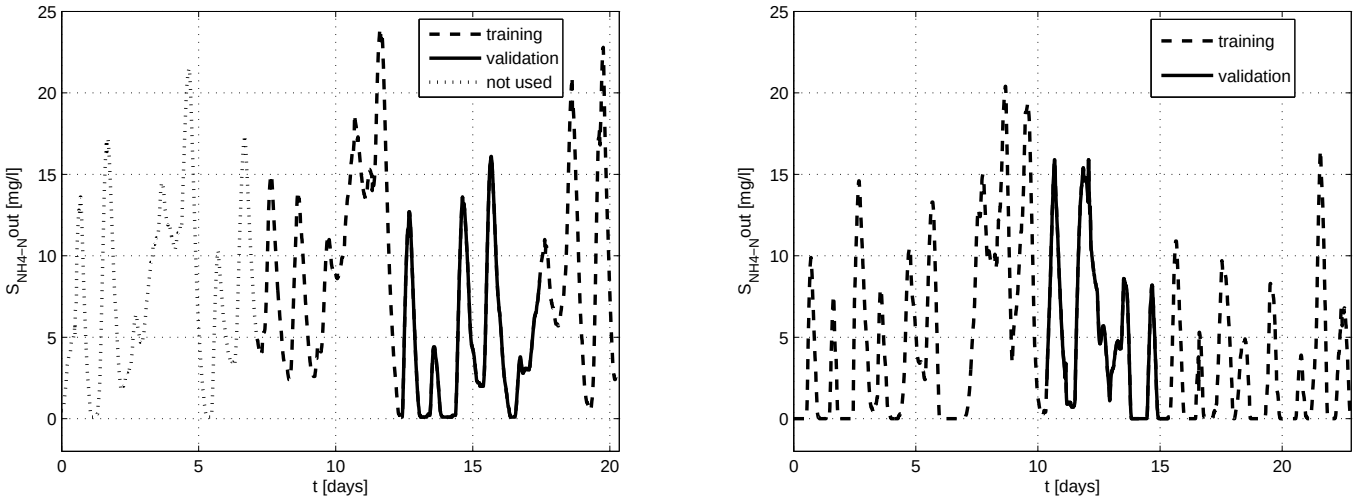
training
validation
not used
S_{NH4-N}out [mg/l]
t [days]
training
validation
S_{NH4-N}out [mg/l]
t [days]

Fig. 7.

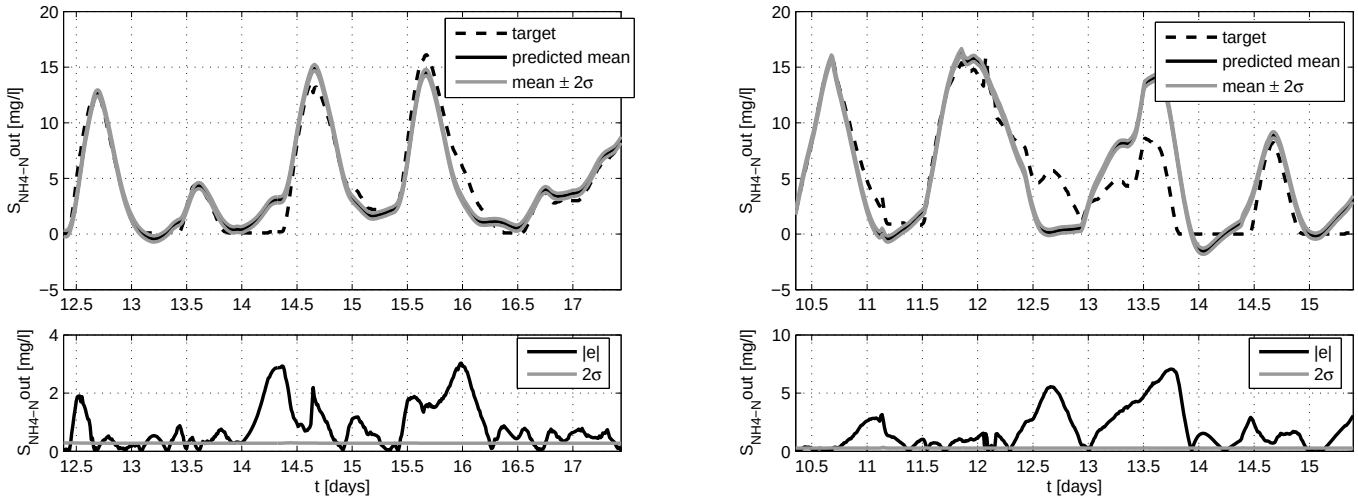

Fig. 8.

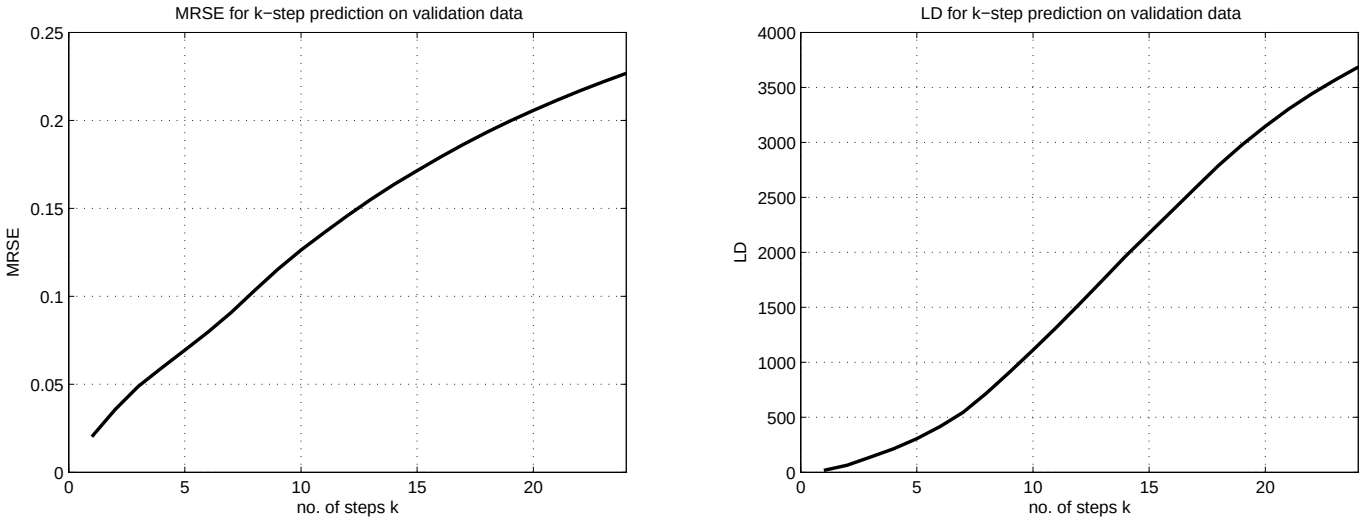
MRSE for k−step prediction on validation data
MRSE
no. of steps k
0
0.05
0.1
0.15
0.2
0.25
0
5
10
15
20
LD for k−step prediction on validation data
LD
no. of steps k
0
500
1000
1500
2000
2500
3000
3500
4000
0
5
10
15
20

Fig. 9.

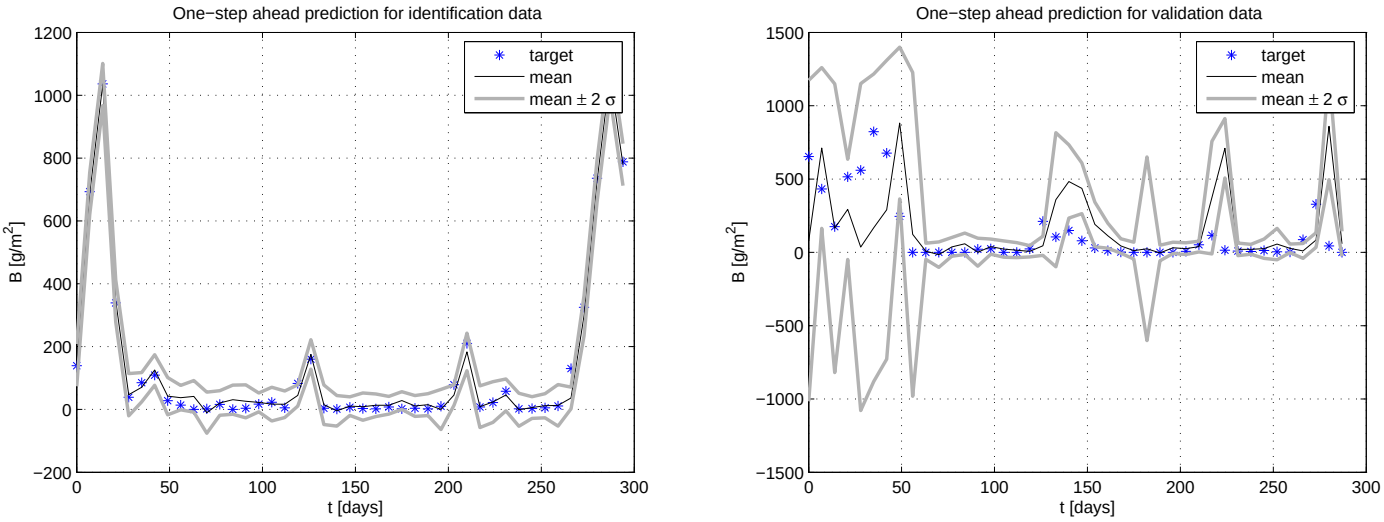
One−step ahead prediction for identification data
target
mean
mean ± 2 σ
B [g/m^2]
t [days]
One−step ahead prediction for validation data
target
mean
mean ± 2 σ
B [g/m^2]
t [days]

Fig. 10.

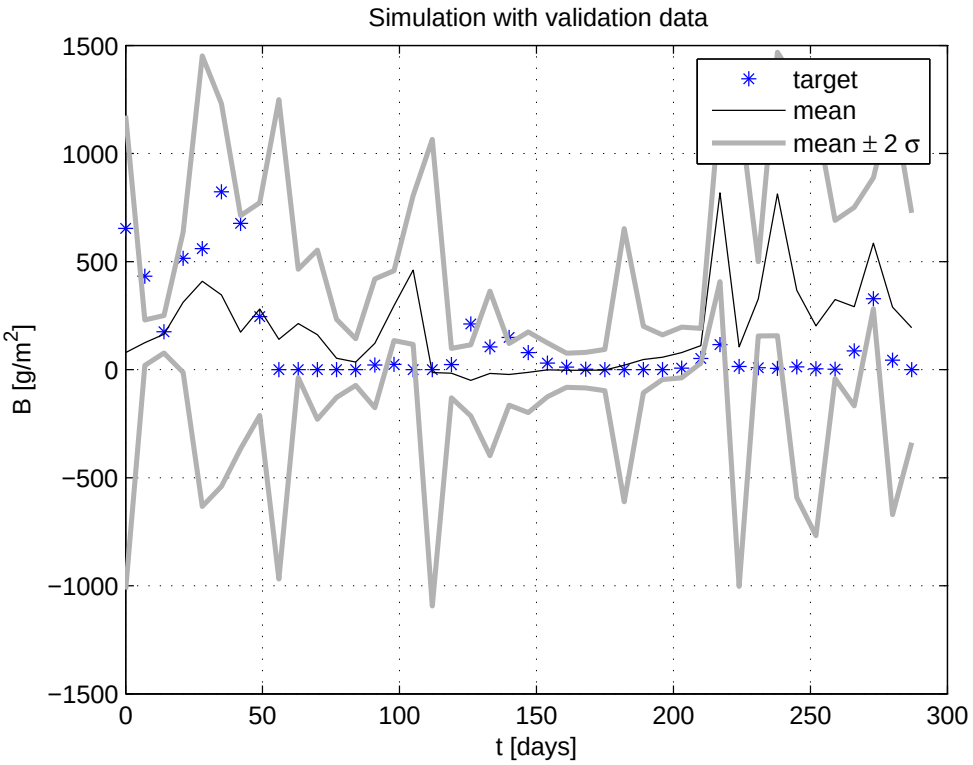
Simulation with validation data
target
mean
mean ± 2 σ
B [g/m^2]
t [days]
1500
1000
500
0
−500
−1000
−1500
0
50
100
150
200
250
300

Fig. 11.